Ensemble Structure of the Modular and Flexible Full-Length Vesicular Stomatitis Virus Phosphoprotein

Cédric Leyrat1, Robert Schneider2, Euripedes A. Ribeiro Jr1, Filip Yabukarski1, Mingxi Yao2, Francine C. A. Gérard1, Malene Ringkjøbing Jensen2, Rob W. H. Ruigrok1, Martin Blackledge2 and Marc Jamin1

1Unit of Virus Host Cell Interactions, UMI 3265 UJF-EMBL-CNRS, 6 rue Jules Horowitz, BP 181, 38042 Grenoble Cedex 9, France
2Protein Dynamics and Flexibility, Institut de Biologie Structurale, UMR 5075 CEA-CNRS-UJF, 41 rue Jules Horowitz, 38027 Grenoble, France

Received 3 February 2012; received in revised form 15 June 2012; accepted 1 July 2012
Available online 9 July 2012

Edited by R. W. Kriwacki

Keywords: phosphoprotein; intrinsically disordered regions; vesicular stomatitis virus; NMR; SAXS

Introduction

Vesicular stomatitis virus (VSV) is the prototype virus of the Rhabdoviridae, a viral family that also includes rabies virus. The genome of these viruses is made of a single-stranded negative-sense RNA molecule. The family Rhabdoviridae is grouped within the order Mononegavirales (MNV) with three other families of non-segmented negative-sense RNA viruses, which also contain important human pathogens, the Paramyxoviridae (measles virus, respiratory syncytial virus, and Nipah virus), the Filoviridae (Ebola virus and Marburg virus) and the Bornaviridae (Borna disease virus). These viruses have different morphologies and different modes of transmission, and they induce different pathologies in their hosts but share similar genomic and structural organizations and similar modes of RNA transcription and replication. The genome of VSV contains five genes that are common to all MNV viruses and are organized in the same order along the RNA genome. They encode successively from the 3′ end to the 5′ end of the nucleoprotein (N), the phosphoprotein (P), the matrix protein (M), the glycoprotein (G) and the large subunit of the RNA-dependent RNA polymerase (L). The G and M...
proteins are involved in the entry of the virus in its host cell and in the assembly and budding of new viral particles, respectively. The three other proteins together with the RNA genome constitute the infectious core of the virus and its transcription/replication machinery. The RNA genome is encased by N, and the N–RNA complex rather than the naked RNA constitutes the actual template for the viral two-subunit polymerase complex comprising the L and P proteins. P is an essential component of the virus that plays multiple roles at different stages of the viral life cycle. First, P acts as a non-catalytic cofactor of the viral polymerase. The L subunit carries out the enzymatic activities required for RNA synthesis, mRNA capping, methylation and poly-adenylation but is unable to bind efficiently to the N–RNA template. P is required for processive RNA synthesis. Since P possesses binding sites for L in its N-terminal part and for the N–RNA template in its C-terminal domain, it is assumed that P correctly positions L on its template and maintains it attached when the polymerase complex moves along the template during transcription and replication. Second, P forms a complex with nascent N named N0. P interacts efficiently to the N–RNA template. On the basis of these studies, a model in which VSV P forms non-globular homodimers was first suggested by functional studies and was recently evidenced by structural studies. The protein contains a long N-terminal intrinsically disordered region (IDR) (IDRNT; amino acids 1–106) and two autonomous folding domains (Pced and PCTD) separated by a disordered linker (amino acids 178–194). The central Pced domain (amino acids 107–177) constitutes the dimerization interface, and the C-terminal PCTD domain (amino acids 195–265) binds to the N–RNA template. Pced was identified by limited proteolysis, and its structure was solved by X-ray crystallography. PCTD was located by a meta-analysis of disorder-predicted the presence of a structured region in the 40 N-terminal amino acids, which contains the binding site for N0. A protein fragment corresponding to the first 60 residues of VSV P (P60) appeared globally disordered in isolation, but two short regions (amino acids 2–12 and amino acids 25–38) transiently adopted α-helical conformations. This N-terminal region constitutes a molecular recognition element (MoRE) that undergoes a disorder-to-order transition upon binding to its physiological partner, N. On the basis of these studies, a model in which VSV P consists of a central dimeric core with two long N-terminal flexible arms and two C-terminal globular domains attached through flexible linkers was proposed. Nevertheless, a description at atomic resolution of the behavior of the entire protein in solution is necessary in order to develop a molecular understanding of the roles played by this protein in the process of viral replication.

Unlike folded proteins, intrinsically disordered proteins and IDRs lack a single stable tertiary structure and are thought to exchange between different conformations in stochastic processes driven by thermal fluctuations. These proteins cannot be described by a unique conformation in solution and are better represented by ensembles of molecular conformations. For multidomain proteins, the atomic structure of isolated well-folded domains can be determined by X-ray crystallography or NMR spectroscopy, while information about the disordered regions and the spatial organization of the domains in the entire protein may be obtained by NMR spectroscopy and small-angle scattering experiments. Various NMR observables, including chemical shifts, scalar couplings, 15N relaxation rates, residual dipolar couplings and paramagnetic relaxation enhancements, provide information about local conformational preferences and long-range order. Small-angle scattering of X-rays or neutrons allows the determination of a limited number of independent parameters that characterize the global size and shape of the molecule in solution and can be used to reconstruct three-dimensional low-resolution models. In combination with the development of computational tools for generating libraries of physically relevant conformations, different methods have been developed over the past few years for building or selecting conformational ensembles on the basis of experimental constraints from NMR, small-angle X-ray scattering (SAXS), or a combination of both methods.

Here, we use NMR spectroscopy and SAXS to characterize the spatial organization of the full-length dimeric VSV phosphoprotein. NMR experiments confirm the existence of folded domains and unfolded regions and, in combination with SAXS data, provide the biophysical basis for an ensemble model of the protein in solution. With this ensemble representation in hand, we discuss the possible roles played by the intrinsic disorder of VSV P in the assembly and functioning of the transcription/replication complex.
and consider the potential advantages conferred by this high level of molecular flexibility.

**Results**

**Full-length VSV phosphoprotein and its fragments**

Full-length VSV P, a short fragment (P<sub>60</sub>) and a long fragment (P<sub>105</sub>) of the IDR<sub>NT</sub> corresponding to the first 60 and 105 residues, P<sub>CED</sub> and P<sub>CTD</sub> were produced as described below (see Materials and Methods). Each protein elutes as a single peak from the size-exclusion chromatography (SEC) column. The molecular mass determined from static light scattering is constant throughout the chromatographic peak, and the polydispersity index (M<subŵ</sub>/M<sub>n</sub>) is lower than 1.01, indicating that each sample is monodisperse.

As previously reported, full-length VSV P elutes from the SEC column as a dimer<br> (Table S1). P<sub>60</sub>, P<sub>105</sub> and P<sub>CTD</sub> are monomeric, whereas P<sub>CED</sub> forms dimers, confirming that this domain contains the dimerization interface (Table S1). In a plot of hydrodynamic radius (R<sub>S</sub>) as a function of molecular mass (Fig. 1), P<sub>CED</sub> and P<sub>CTD</sub> behave similar to folded globular proteins, while P<sub>60</sub> and P<sub>105</sub> behave similar to unfolded proteins. Full-length P appears at an intermediate position in accordance with the presence of both folded domains and IDRs in similar proportions.

**NMR spectroscopy**

Full-length VSV P and its domains were investigated by NMR (Fig. 2). The<br>£H–15N two-dimensional heteronuclear single quantum coherence (HSQC) spectrum of P<sub>105</sub> exhibits poor chemical shift dispersion of the amide £H resonances (Fig. 2b) as is typical for a disordered protein. Chemical shifts depend on the backbone ϕ/ψ dihedral angles, and in disordered systems, they are highly sensitive to the presence of transient secondary structure. The SSP score for P<sub>105</sub> reveals the presence of two transient α-helices in the N-terminal moiety of the peptide (amino acids 2–16 and amino acids 25–31), in agreement with data previously recorded for P<sub>59</sub> (Fig. 2a) and with the consensus predictions of disorder (Fig. S1a) and secondary structure (Fig. S1b).

The NMR spectra of isolated P<sub>CED</sub> (amino acids 107–177) and P<sub>CTD</sub> (amino acids 194–265) confirm that both domains are folded in solution. The HSQC spectra of these fragments show well-dispersed amide £H resonances, characteristic of folded proteins (Fig. 2c and d). P<sub>CED</sub> (79 residues including the eight-amino-acid histidine tag), the HSQC spectrum exhibits 68 clearly identifiable signals. For P<sub>CTD</sub>, the backbone amide and side-chain resonances were assigned in a previous study and used to generate a high-resolution structure of this domain.

The HSQC spectrum of the full-length P dimer contains two sets of resonances of different intensities. A set of intense signals with limited chemical shift dispersion corresponds well to the spectrum recorded for P<sub>105</sub> (Fig. 2e), indicating that the N-terminal part is flexible with the same conformational behavior as the isolated peptide and negligible interdomain contacts. Some additional sharp resonances that may correspond to residues in the flexible linker between P<sub>CED</sub> and P<sub>CTD</sub> are visible in the spectrum of full-length P. A second set of broader and less intense resonances corresponds to resonances found in the spectra of isolated P<sub>CED</sub> and P<sub>CTD</sub>, showing that these two regions remain structured in the context of the full-length protein (Fig. 2f and g). The presence of long IDRs in VSV P exerts a frictional drag that slows down the overall tumbling rate of the molecule and may explain the decrease of signal intensity observed for resonances in the globular domains as compared to the spectra of the isolated domains (Fig. 2c and d).

These results clearly indicate that (1) the N-terminal region of P exchanges between multiple conformers on a fast timescale and contains transient α-helical elements, that (2) P<sub>CED</sub> and P<sub>CTD</sub> are folded in the full-length protein and that (3) there are no significant interactions between the different regions of the molecule. These observations form the basis for subsequent ensemble modeling approaches.
The shapes of the SAXS curves (Fig. 3a) and of the Guinier plots (Fig. 3b) obtained for full-length P are independent of protein concentration, indicating the absence of significant aggregation. The model-free analysis of small-angle scattering curves provides information about the ensemble-averaged dimensions and shape of the scattering particles.\(^6^4\) The radius of gyration, \(R_g\), determined from the Guinier approximation is 5.4±0.1 nm in the range \(Q\cdot R_g < 1.3\) and is 5.8±0.4 in the range \(Q\cdot R_g < 1.0\) (Table S2), in agreement with the value of 5.6±0.1 nm determined from the pair distribution function \(P(r)\) (Fig. 3c) and with the value of 5.3±0.1 nm measured previously by small-angle neutron scattering.\(^1^9\) This \(R_g\) value is significantly larger than that expected for a globular protein of the same.

**Small-angle X-ray scattering**

The shapes of the SAXS curves (Fig. 3a) and of the Guinier plots (Fig. 3b) obtained for full-length P are independent of protein concentration, indicating the absence of significant aggregation. The model-free analysis of small-angle scattering curves provides information about the ensemble averaged dimensions and shape of the scattering particles.\(^6^4\) The radius of gyration, \(R_g\), determined from the Guinier approximation is 5.4±0.1 nm in the range \(Q\cdot R_g < 1.3\) and is 5.8±0.4 in the range \(Q\cdot R_g < 1.0\) (Table S2), in agreement with the value of 5.6±0.1 nm determined from the pair distribution function \(P(r)\) (Fig. 3c) and with the value of 5.3±0.1 nm measured previously by small-angle neutron scattering.\(^1^9\) This \(R_g\) value is significantly larger than that expected for a globular protein of the same.
molecular mass, in agreement with the presence of more extended conformations expected for a protein with extensive disorder. For elongated or disordered proteins, the linearity of the Guinier plot is restricted to a lower range of $Q$ values than for globular proteins, and for a random-coil chain, it is recommended to use the Debye plot rather than the Guinier plot because its linearity extends to a larger range of $Q$ values. For VSV P, the linearity of this plot is restricted to a narrow range of $Q \cdot R_g$ values ($0.5 < Q < 3.0$) and yields a value of $6.0 \pm 0.1$ (Fig. S2a and Table S2). These two approaches, thus, set limits on the estimation of the $R_g$ value.

The pair distribution function obtained by Fourier transformation of the experimental curve exhibits an asymmetrical shape with a large $D_{max}$ value (Fig. 3c), whereas the Kratky plot presents a maximum at 0.5 nm$^{-1}$ and a flat region above 2 nm$^{-1}$ (Fig. 3d), also supporting the presence of folded domains and disordered regions. However, despite the presence of several folded domains, both curves are smooth, lacking interdomain correlation peaks as would be expected for a multidomain protein in which the folded domains are positioned at fixed distances from each other. The differences observed in the pair distribution functions (Fig. 3c) and Kratky plots (Fig. 3d) for the three protein concentrations arise from minor variations within the limits of experimental errors (Fig. S2b).

Bead models generated with the program DAMMIN by ab initio reconstruction from the SAXS data appear as elongated cylinders (Fig. S3a and b) that could accommodate the structures of one P$_{CED}$ dimer and two P$_{CTD}$ domains (Fig. S3c). Models of similar shape were obtained with the program BUNCH that combines atomic structures for P$_{CED}$ and P$_{CTD}$ and bead models for the unfolded strands (Fig. S3d and e). Such models may be considered as population-weighted representations of a potentially highly heterogeneous conformational ensemble. However, these programs, designed to model compact structures, are constrained by the large $D_{max}$ value; therefore, modeling of highly disordered molecules such as VSV P as a single
Fig. 4 (legend on next page)
conformer is biased toward elongated shapes that may not necessarily reflect the actual structural organization of the system.\textsuperscript{33,67} It should be noted, however, that modeling with DAMMIN or BUNCH yielded good fits to the scattering data with $\chi$ values of 0.056 and 0.293, respectively. Therefore, it must be stressed that the goodness of fit of the SAXS data cannot be used as the only criterion for discriminating between different possible structural models.\textsuperscript{48}

**Modeling VSV P as an ensemble of conformations**

To account for the mobility of the polypeptide chain observed by NMR spectroscopy, we modeled the structure of the P dimer as an ensemble of conformers rather than as a single structure. An ensemble of conformations of the VSV P dimer was built with the program flexible-meccano.\textsuperscript{46,68} This program generates physically relevant conformers of disordered proteins and uses a simple exclusion procedure to avoid clashes within the protein. For modeling the IDR\textsubscript{NT} region of P, we used a database of backbone dihedral angles compiled by the ensemble selection algorithm ASTERIODS\textsuperscript{70} in order to match predicted $^{13}$C\textsubscript{α}, $^{13}$C\textsubscript{β} and $^{13}$C\textsubscript{γ} chemical shifts for this region of P (calculated using the program SPARTA\textsuperscript{69}) with experimental values measured on P\textsubscript{105} (Fig. 4a and b; see Materials and Methods). The dimeric P\textsubscript{CED} and the two P\textsubscript{CTD} domains were constructed in their folded forms. In this way, an initial ensemble of 8000 conformers was built with flexible-meccano. This ensemble exhibits broad Gaussian distributions of $R_g$ and $D_{max}$ values as expected for a polypeptide chain obeying random-coil statistics (Fig. 4c–f); however, the average SAXS curve calculated for the 8000 models using CRYSOl\textsuperscript{70} does not fit the experimental curve ($\chi = 0.75$) (Fig. S4).

From this initial pool, sub-ensembles containing different numbers of conformers were selected with the program GAJOE (Genetic Algorithm Judging Optimization of Ensembles) for which the average theoretical curve reproduces the experimental SAXS curve up to $Q = 2.0$ nm\textsuperscript{−1} (Fig. 4g and Fig. S4).\textsuperscript{53} The $\chi$ value\textsuperscript{86} [Eq. (2)] used here to evaluate the quality of the fit decreases with increasing number of models in the selected ensemble and levels off for selected ensembles of more than five models (Fig. 4h). This result clearly shows that the SAXS curve is better represented by an ensemble of conformers than by a single conformer. Also, similar fits were obtained when the initial ensemble was reduced to 2000 conformers, suggesting that 8000 conformers are sufficient to reproduce the conformational diversity of the molecule at the resolution level accessible from the SAXS curve (Fig. 4i).

Selected ensembles of 50 conformers adequately reproduce the SAXS curve (Fig. 4g) and the NMR chemical shifts for the 105 N-terminal residues (Fig. 4a and b), providing an atomic representation of VSV P in agreement with both sets of experimental data. The different conformers obtained in independent selection processes exhibit a broad range of $R_g$ and $D_{max}$ values, and independent rounds of selection yielded similar results (Fig. S5 and Table S3). Figure 4c and d show the distributions of $R_g$ and $D_{max}$ for a representative ensemble of 50 conformers, and Fig. 4e and f shows the distributions of $R_g$ and $D_{max}$ for a representative minimal ensemble of 5 conformers. The averaged $R_g$ and $D_{max}$ values of the selected ensembles are within the limits of values determined from the experimental curves (Table S3) but are significantly larger than those of the initial ensemble (mean $R_g$ value of 5.1 nm with a standard deviation of 0.6 nm and mean $D_{max}$ value of 17 nm with a standard deviation of 3 nm). The distribution functions of the selected ensembles are shifted to higher $R_g$ and $D_{max}$ values as compared to those of the initial ensembles (Fig. 4b–e). This suggests that P is more extended

---

**Fig. 4.** Modeling of VSV P as an ensemble of conformers with GAJOE. (a) Experimental and calculated $C^\alpha$ chemical shifts for VSV P ensembles. The panel shows $C^\alpha$ secondary chemical shifts. The black line shows experimental chemical shifts for P\textsubscript{105} and the red and blue lines show chemical shifts calculated using SPARTA for IDR\textsubscript{NT} residues of the initial 8000-strong full-length VSV P ensemble (red) and of a selected ensemble of 50 conformers (blue). (b) Experimental and calculated $C^\beta$ chemical shifts for VSV P ensembles. (c) $R_g$ distributions for an ensemble of 50 conformers. The black curve shows the $R_g$ distribution calculated for the initial pool of 8000 conformers. The red curve shows the $R_g$ distribution for the selected ensemble of 50 conformers that fits the experimental SAXS data. Typically, the selection process performed by GAJOE involves 50 successive cycles, and the distribution shown in red corresponds to the distribution of $R_g$ values for the models selected in the best ensemble of each of these successive selections. The black bars show the $R_g$ values for the 50 conformers that form the best of all of these ensembles. (d) $D_{max}$ distribution for an ensemble of 50 conformers. The same color scheme as in (a) is used for the $D_{max}$ distributions. (e) $R_g$ distributions for an ensemble of five conformers. (f) $D_{max}$ distribution for an ensemble of five conformers. (g) Fitting of SAXS data. The experimental SAXS curve is shown in black, and the fitted line for a selected ensemble of 50 conformers is shown in red. The $\chi$ value is 0.207. The lower panel shows the residuals of the fit. (h) Variation of the goodness of the fit ($\chi$) with the number of models in the selected ensemble. The selection with GAJOE was repeated for ensembles containing 1–50 conformers. The black circles are for selections performed without allowing multiple selections of the same conformer, whereas white circles are for selections allowing multiple selections. (i) Variation of the goodness of the fit ($\chi$) with the number of models in the initial ensemble. $\chi$ values are for selection of ensembles of five conformers without allowing multiple selections of the same conformer.
than expected based on pure random-coil sampling of the conformational space assigned to the flexible regions. The IDRNT is rich in acidic residues; thus, charge repulsion most likely accounts for this more extended organization. The alignment of the different models from one selected ensemble obtained by superimposing their PCTD reveals that the P dimer samples a large conformational space in solution (Fig. 5), and the comparison of different ensembles indicates that this behavior is independent of the number of models in the selected ensemble (Figs. S5 and S6).

The next questions are to test whether a discrete number of conformers are sufficient for representing the conformational diversity of such a flexible molecule and whether fitting the experimental SAXS curve with an ensemble of conformers allows one to discriminate between different models of VSV P. To answer these questions, we performed simulations with reference ensembles. In this approach, a set of conformers is chosen from the first pool of 8000 conformers to constitute the reference ensemble, and a synthetic scattering profile is calculated for this reference ensemble using CRYSOl. Then, the program GAJOE is used to select a subset of 1–50 conformers from the initial pool of conformers that minimize the χ value between the scattering curves calculated for this selected subset and those calculated for the reference ensemble (Fig. S7). In the first case considered, the reference ensemble consisted of a single conformer arbitrarily chosen from the initial pool of models. On the basis of its theoretical scattering curve, GAJOE is able to correctly retrieve this same conformer from the initial pool (selected subset of one model). Attempting to represent this unique conformer with an ensemble of several conformers decreases the quality of the fit, as judged by the χ value, with increasing numbers of conformers. When multiple selection of the same model is allowed, the program GAJOE selects the same model several times for ensembles containing up to five models. This clearly indicates that the system is best represented by a unique conformer. In the second case, the reference ensemble consisted of a set of 10 different conformers. Here, the quality of the fit significantly improves with increasing number of models in the selected ensemble up to ensembles of 10 conformers. In a selected ensemble of 20 structures, none of the models selected by GAJOE corresponds to the models constituting the reference ensemble, and the lowest r.m.s.d. value between conformers of the selected and reference ensembles is 2.3 nm (r.m.s.d. values ranging between 2.3 and 5.9 nm) (Fig. S8). Thus, in this case, the conformational diversity of the reference system is best represented by an ensemble of conformers. In the third case, the reference ensemble consisted of 8000 structures. Again, the quality of the fit improves with increasing number of models in the selected ensemble but, as in the second example, exhibits little improvement above 10 conformers. Thus, in line with previous comparisons applied to completely disordered systems and multidomain proteins, these simulations demonstrate that, in the case of proteins containing folded and disordered domains, selection using...
GAJOE clearly differentiates a situation where the protein adopts a unique structure from one where it is averaged over multiple conformers. However, because of the limited information content of the SAXS curve, discrete ensembles of 10 or 20 conformers appear to be sufficient for describing such a system of large conformational diversity.

### Discussion

Characterizing the overall structure of multidomain proteins that contain folded domains concatenated with long IDRs is complicated by the intrinsic conformational heterogeneity of the disordered segments. The size of the VSV P particle determined by SEC and SAXS indicates a non-globular character suggesting either an elongated molecule or an ensemble of rapidly interconverting conformers. The smooth profiles of the pair distance distribution function and Kratky plot derived from the SAXS data (Fig. 3c and d) could be explained either by a physically unrealistic organization in which the folded domains adopt fixed positions in space or, more likely, by conformational averaging. However, definite evidence of the conformational heterogeneity of VSV P emerged from NMR spectroscopy. The NMR data clearly confirm the presence of disordered regions in the full-length protein and show that the disordered and ordered regions behave independently. NMR is particularly suitable for studying the dynamic behavior of disordered polypeptide chains, providing amino-acid-specific identification of transient structural elements, motional correlations and long-range contacts in globally disordered proteins. For VSV P, NMR clearly shows that, in the context of the full-length P dimer, the N-terminal region of each P monomer interconverts between multiple conformations on a timescale faster than the millisecond range, while PCEF and PCTD adopt the stable, folded structures that were determined from isolated domains. Although the slow tumbling of the full-length dimeric protein leads to resonance broadening and low signal intensity for the folded domains, nevertheless, we were able to exclude significant interactions between unfolded and folded domains. Only one set of resonances is observed in the NMR spectra, demonstrating that the chains of the VSV P dimer behave identically. We have incorporated this information into an ensemble description of the molecular behavior of VSV P that conveys the dynamic character of the molecule and its departure from pure random-coil behavior. The different models of the selected ensemble represent different possible conformations of VSV P that collectively reproduce the experimental scattering curve, as well as the local conformational behavior and chemical shifts of the N-terminal disordered region as identified from NMR. In the selected ensembles, P dimers adopt a large variety of conformations with different dimensions and relative orientations between the N- and C-terminal arms. It should also be noted that there is not a unique subset of conformers that reproduce both SAXS and NMR data. The conformers of one selected ensemble are, thus, merely examples of the structure of the protein that are in agreement together with the experimental data.

### Comparison with the phosphoprotein of other MNV viruses

VSV P shares a similar modular organization not only with the phosphoprotein of rabies virus, another member of the *Rhabdoviridae*, but also with the phosphoprotein of different viruses from the *Paramyxoviridae* and the *Bornaviridae*. In all these proteins, long IDRs alternate with structured domains, and similar regions of the protein carry out similar functions during the replication cycle. The N-terminal region is involved in binding RNA-free N, the central region of the protein that contains a homo-oligomerization domain, while the C-terminus binds the N–RNA template. All these proteins also interact with the L subunit of the RNA-dependent RNA polymerase, although the site of interaction is not clearly mapped, and specific functions or post-translational modifications have been located either in the IDRs or in the folded domains. Thus, the model proposed here for VSV P, in which N-terminal and C-terminal flexible extensions are tethered to a central homo-oligomeric core, may also apply to the P proteins from these other MNV viruses. Although evidence that the P protein contains disordered and ordered regions have been gathered for different viruses, this study provides the first structural description of a full-length oligomeric phosphoprotein obtained on the basis of experimental data.

### Importance of size and flexibility of VSV P for its functions

VSV P plays multiple roles and interacts with different viral and cellular partners throughout the viral life cycle. Here, we endorse the view that the remarkable dimensions and flexibility of the protein dimer have important implications in the assembly of the virus and the functioning of the replication machinery. By sampling a large volume space, the VSV P dimer may increase the probability of encountering its partners in solution and, therefore, increase the on-rate for binding as proposed in the “fly casting” mechanism (Fig. 6). It is noticeable that the long N-terminal flexible arm of VSV P contains binding sites for multiple partners, including phosphorylation sites, a nuclear export signal, a
MoRE for binding $N^0$ and a binding site for the L protein. A main function of P is to position the polymerase on the N–RNA complex and to keep it attached when the polymerase moves along the template, in particular, when it pauses at the intergenic regions during transcription. When P is bound to the N–RNA template in the infected cell, the N-terminal arms extend into solution and may serve to catch L and position it onto the template. Also, it has been proposed that P moves by cartwheeling along the N–RNA template, carrying the L subunit. This model was proposed on the basis of results obtained with Sendai virus (Paramyxoviridae) in which oligomerization of the protein is indispensable. In contrast, the homodimeric domain of rabies virus P is dispensable for transcription, and we propose an alternative mechanism for the progression of the polymerase of rhabdoviruses along its template. In this new model, P remains attached to the N–RNA template through its $P_{\text{CTD}}$. Currently, it is not clear whether both $P_{\text{CTD}}$ domains of a P dimer can bind simultaneously to the N–RNA complex, but our ensemble model shows that it is physically possible. In the 8000 model ensembles, which gives an idea of the physically accessible conformations of this protein, $P_{\text{CTD}}$ can project up to 9 nm away from $P_{\text{CTD}}$, and thus, it is possible that both $P_{\text{CTD}}$ domains bind simultaneously to neighboring sites of the N–RNA complex.

By sampling a large conformational space around their attachment points on the N–RNA template, P dimers may act similar to polymer brushes by entropic exclusion, as demonstrated for disordered regions of nucleoporins, neurofilaments or microtubule-associated proteins. Once a P molecule is bound to the N–RNA complex through its C-terminal domains, its thermally driven motions may restrict access to the space around this molecule to other P molecules and thereby regulate the spacing between P molecules along the N–RNA complex (Fig. 7a). Entropic brushes have been
involved in controlling the spacing between neurofilaments or microtubules. The estimated number of P and N proteins in the VSV virion results in one P dimer bound to every five N protomers. If P molecules are distributed at regular intervals along the template, the capture radii of the N-terminal arms of adjacent P molecules overlap (Fig. 6), and we propose that the polymerase moves along the template by jumping from one P to the next (Fig. 7b). Also, in the generated ensemble of 8000 conformers, the distance separating the center of mass of PCTD from that of IDRNT in dimeric P can extend up to 18 nm, indicating that once P is bound to the N–RNA complex through its C-terminal domain, the search volume accessible to the N-terminal extremity of each monomer extends to sites located further than five N protomers from the binding site (Fig. 6). The long N-terminal arm of P could fetch the polymerase upstream, keep it attached to the template during the time it replicates a stretch of RNA encapsidated by roughly five N protomers (~45 nt), and then deliver it downstream to the next P molecule. A mechanism of folding upon binding may provide a means of specific recognition of P by L without the corollary of high affinity and may allow the L protein to exchange between bound P proteins. It now remains to be determined if the L protein possesses a single binding site for P that would imply that the incoming N-terminal arm exchanges with the previously bound one or if it possesses multiple binding sites that would allow the simultaneous attachment of multiple P N-terminal arms.

The dimensions and plasticity of P may also serve in the assembly of new N–RNA complexes. The encapsidation of the newly synthesized RNA genome requires the delivery of soluble RNA-free N0 to the site of RNA synthesis. The N0–P complex can bind to the nucleocapsid but little is known about the reaction by which N0 is transferred from P to the RNA. When an N0–P complex is bound to the N–RNA template through the C-terminal domain of P, the flexibility of the P molecule may correctly position and orient the N0 molecule within the replication complex and deliver it to the site of encapsidation.

Finally, it is worth noting that the dimensions of the phosphoprotein are remarkable as compared to the dimensions of the virion. The VSV particle has a bullet shape of about 200 nm in length and 70 nm in diameter. In this structure, the N–RNA complex curls into a helix that constitutes a tubular inner leaflet and defines an internal cavity of length 150 nm and a radius of 15 nm. The average \( D_{\text{max}} \) of the P dimer in solution of 18 nm is longer than the radius of this inner cavity and about one-eighth of its length. However, in accordance with the partial specific volume of proteins, this cavity is large enough to accommodate 225 molecules of P and 55 molecules of L that are present in the virion, as well as 500 M proteins that are also present in virion but are not involved in the intermediate shell of the particle. It is not known how these proteins are packed within the cavity, but because the C-terminal domain of N that binds the C-terminal domain of P is oriented toward the internal cavity, it is possible...
that P is bound to the N–RNA template. Also, since the rise of the N–RNA helix is 5.1 nm, \( P_{\text{CTD}} \) domains of one P dimer could bind to successive rungs of the N–RNA helix and stabilize the virion.

### Materials and Methods

#### Sample preparation

VSV full-length P, \( P_{60} \) and \( P_{\text{CTD}} \) were produced and purified as previously described, \(^{19,26,28} \) and similar procedures were set up for purifying \( P_{105} \) and \( P_{\text{CED}} \). The homogeneity of each sample was checked by SEC combined with detection by multi-angle laser light scattering and refractometry. \(^{19,29} \) SEC was performed with an S200 Superdex column (GE Healthcare) equilibrated with 20 mM Tris–HCl at pH 7.5 containing 150 mM NaCl. Separation was performed at 20 °C with a flow rate of 0.5 ml min \(^{-1} \). We injected 50 μl of a protein solution at a concentration of 5.5 mg ml \(^{-1} \). Online multi-angle laser light scattering detection was performed with a DAWN-EOS detector (Wyatt Technology Corp., Santa Barbara, CA) using a laser emitting at 690 nm. Data were analyzed, and absolute molecular weights (\( M_w \)) were calculated using the ASTRA software (Wyatt Technology Corp.) as described previously by Gérard et al. \(^{33} \) The S200 Superdex column was calibrated with proteins of known Stokes’ radii (\( R_S \)): \(^{35} \) chymotrypsinogen (\( R_S =2.3 \) nm), RNase A (\( R_S =1.9 \) nm), ovalbumin (\( R_S =3.0 \) nm), albumin (\( R_S =3.4 \) nm), aldolase (\( R_S =4.7 \) nm), catalase (\( R_S =5.2 \) nm), ferritin (\( R_S =6.8 \) nm) and thyroglobulin (\( R_S =7.9 \) nm).

#### NMR spectroscopy

NMR experiments were performed on Varian spectrometers operating at \(^1{H} \) frequencies of 600 and 800 MHz. \(^{15} \text{N} \) HSQC spectra of VSV \( P_{105} \) and full-length P shown in Fig. 2 were recorded on samples in 20 mM Tris–HCl, 150 mM NaCl, 50 mM Glu and 50 mM Arg with 10% D\(_2\)O at 20 °C and pH 7.5 at a 600-MHz \(^1{H} \) resonance frequency. Resonance assignment of VSV \( P_{105} \) was carried out using a double-labeled \((^{15}\text{N},^{13}\text{C})\) sample of the peptide with the buffer adjusted to pH 6.0 to avoid loss of resonances due to fast amide proton exchange with the solvent protons. Assignment experiments were carried out at 25 °C and a \(^1{H} \) resonance frequency of 600 MHz. The assignment was obtained from a series of BEST-type triple-resonance experiments: \(^{36} \text{HNCO, HN(CO)}\text{CA, HN(CO)}\text{CA, HN(CA)}\text{CA, HN(COCA)}\text{CB, and HN(CA)}\text{CB.} \) The \(^{1} \text{H} \) \(^{15} \text{N} \) HSQC spectrum of the central \( P_{\text{CTD}} \) domain was recorded at 20 °C and pH 7.5 at a \(^1{H} \) resonance frequency of 600 MHz, using a buffer containing 20 mM Hepes, 150 mM NaCl and 1 mM DTT. The spectrum of the C-terminal \( P_{\text{CTD}} \) domain in Fig. 2 was recorded at a \(^1{H} \) resonance frequency of 800 MHz, at 20 °C and pH 7.5 in 20 mM Tris–HCl buffer containing 150 mM NaCl. All spectra were processed with NMRPipe\(^{37} \) and analyzed using SPARKY.\(^{38} \) Automatic assignment of \( P_{105} \) resonances on the basis of SPARKY peak lists was performed using the program Mars.\(^{39} \)

### SAXS experiments

The scattered intensity of full-length VSV P was collected for scattering vector values (\( Q =4\pi \sin(\theta)/\lambda \)) ranging from 0.1 to 3.0 nm \(^{-1} \) over a concentration range of 1–11 mg ml \(^{-1} \). SAXS data were collected at the European Synchrotron Radiation Facility (Grenoble, France) on beamline ID14-3. The sample-to-detector distance was 1 m, and the wavelength of the X-rays was 0.0995 nm. Samples were contained in a 1.9-mm-wide quartz capillary. The time of exposure was optimized for reducing radiation damage. Data acquisition was performed at 20 °C.

Data reduction was performed using the established procedure available at ID14-3, and buffer background runs were subtracted from sample runs. The radius of gyration and forward intensity at zero angle \( I(0) \) were determined with the program PRIMUS.\(^{100} \) The maximum dimension \( (D_{\text{max}}) \) value was adjusted such that the \( R_s \) value obtained from GNOM agreed with that obtained from the Guinier analysis.

#### Modeling of VSV P

##### Modeling VSV P as a single conformer

The program DAMMIN was used to generate \( ab \) \textit{initio} low-resolution models filled with spheres (dummy atoms) by fitting the calculated scattering curve to the experimental curve. \(^{39,66} \) This program uses a simulated annealing minimization procedure to find the optimum positions of the dummy atoms, starting from a spherical volume of defined radius. By imposing connectivity constraints, DAMMIN minimizes the interfacial area between the molecule and the solvent and therefore favors compact models. We performed 20 independent runs of DAMMIN with no symmetry restriction.

The program BUNCH\(^{66} \) was used to perform rigid-body modeling of the full-length protein using the known structures of VSV \( P_{\text{CED}} \) and \( P_{\text{CTD}} \). The regions of the protein for which no atomic structure was available were modeled by dummy residues. By a simulated annealing protocol, the optimal position and orientation of the folded domains and the conformation of the flexible parts were obtained by fitting the calculated scattering curve to the experimental curve. We performed 16 independent runs of BUNCH with a 2-fold symmetry restriction.

##### Modeling VSV P as an ensemble

An ensemble of 8000 full-length VSV P conformers was generated with the program flexible-meccano.\(^{46} \) The models consisted of P dimers and included the high-resolution three-dimensional structures of the central
and C-terminal domains.\textsuperscript{23,26} To accurately model the disordered N-terminal region, we first applied the ASTEROIDS selection algorithm\textsuperscript{50} to the P\textsubscript{105} fragment in order to obtain an ensemble of conformers of this region in agreement with experimental $^{13}$C\textsubscript{a}, $^{13}$C\textsubscript{b} and $^{13}$C NMR chemical shifts. Briefly, in successive rounds of selection, 5 × 200 pairs of ($\psi, \phi$) dihedral angles were selected for each residue from a pool of 10,000 P\textsubscript{105} conformers assembled by flexible-mecano such that the difference between $^{13}$C chemical shifts predicted by the program SPARTA\textsuperscript{69} and corresponding experimental values was minimized. The selected dihedral angles were then used in generation of a new set of 10,000 conformers for the next round of selection. The ($\psi, \phi$) database thus obtained after four rounds of selection was then used by flexible-mecano to build the IDR\textsubscript{NT} of full-length VSV P. Residues in the linker region between P\textsubscript{CED} and P\textsubscript{CTD} were stochastically assigned amino-acid-specific backbone dihedral angles in agreement with random-coil statistics. In the disordered regions, side chains were constructed using the program SCCOMP.\textsuperscript{102} Scattering curves were calculated with CRYSOL,\textsuperscript{70} and selections were performed with the program GAJOE.\textsuperscript{83}

We selected optimized ensembles comprising 1–50 conformers that minimize the discrepancy between the experimental and calculated curves according to the following equation:

$$x^2 = \frac{1}{K-1} \sum_{j=1}^{K} \left( \frac{\mu(L(Q_j)) - L_{exp}(Q_j)}{\sigma(Q_j)} \right)^2$$

where $K$ is the number of points in the experimental curve, $\sigma$ is the standard deviation and $\mu$ is a scaling factor. Both for the initial ensemble of 8000 conformers and for selected ensembles of 50 conformers, NMR chemical shifts were back-calculated using the program SPARTA\textsuperscript{69} that was already used in derivation of the backbone dihedral angle database for IDR\textsubscript{NT} of P.

Acknowledgements

This work was supported by grants from the French National Research Agency (ANR) [ANR-07-001-01 (ANRAGE) and ANR JCJC 2010 Protein Disorder (M.R.J.)], the FINOVI foundation and Lyonbiopôle. We thank the Fédération TGIR RMN Très Hauts Champs for providing access to the 800-MHz NMR spectrometer in Grenoble. We thank Dr. G. Zaccai (Institut Laue–Langevin, Grenoble) for helpful discussions. R.S. was supported by the Human Frontier Science Program (long-term fellowship LT000322/2011-L). E.A.R. was supported by postdoctoral fellowships from both ANR and FINOVI programs. C.L. and F.G. were supported by MENRT fellowships from the French government. We thank the Partnership for Structural Biology for the excellent structural biology environment.

Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2012.07.003

References


Solution Structure of the VSV Phosphoprotein


